I/S: A JOURNAL OF LAW AND POLICY FOR THE INFORMATION SOCIETY

The Big ESI: Going from Big to Better in E-discovery

BETSY BARRY*, SUZANNE E. SMITH†,
BETH-ANNE SCHUELKE-LEECH‡, CLAYTON DARWIN§

I. INTRODUCTION

The sheer volume of information generated by people and devices in both personal and professional contexts is growing at an

* Betsy Barry, PhD, Is the VP for Research and Development at Illocution Inc. She earned her PhD from the University of Georgia in Linguistics. For the past decade, Dr. Barry has worked as a forensic linguist doing large-scale language investigation in e-discovery for civil litigation. She is currently a visiting scholar in the Engineering Department at the Ohio State University doing big data policy research on safety and innovation in the automotive industry.

† Suzanne E. Smith, J.D., is a founder of Illocution Inc. She is a graduate of The American University and earned her J.D. from Santa Clara University School of Law. Ms. Smith helps clients develop data driven strategies for e-discovery and academic research. She specializes in performing focused, linguistic based investigations of large corpora with particular experience in pharmaceutical and medical device products liability litigation. She is the CEO of Illocution Inc.

‡ Beth Anne Schuelke-Leech, PhD, is an Assistant Professor in the John Glenn School of Public Affairs, The Ohio State University. Her research is focused on innovation policy and the connections between engineering, business, finance, and policy. She is a Senior Research Fellow of Policy Analytics at Illocution Inc., looking at applying text and linguistic analysis of large policy data collections.

§ Clayton Darwin, PhD, is the Chief Technology Officer and Senior Developer at Illocution Inc. He holds a doctorate in Linguistics from the University of Georgia. For the past decade Clayton has worked in the legal field in e-discovery and litigation support. His primary expertise lies in the development of computer-assisted methods and applications for large-scale document analysis and investigation.

accelerated pace.[1] Much of the conversation surrounding this ever-expanding universe of information has necessarily focused on size and exigencies related to managing and assessing data on a grand scale.[2] Nowhere has this been more evident than in the legal profession with respect to electronic discovery ("e-discovery").[3] In the era of big data, the duty to preserve and produce electronically stored information qua evidence has prompted a myriad of issues in e-discovery, particularly in the context of large civil suits.[4] Issues concerning the quantity of electronically stored information ("ESI") tend to overwhelm and obfuscate equally important matters, matters that go beyond scope and breadth, but that get to the heart of any data-centric, empirical research – the most important being matters of data quality. By now, e-discovery and the legal profession in general, have had over a decade to acclimate to growing quantities of data, from disparate sources, all housing valuable information that warrants accessing and investigating. It is time to shift the focus of the conversation from size or *quantity* of data, to the *quality* of data, or from big data to better data, or even best data, as it were. When it comes to data, quality eclipses quantity.

This article examines big data in the context of e-discovery processes, from identification and collection to production and investigation, while also looking at the characteristics of discoverable ESI and discussing how they affect the quality of data with respect to these processes. It examines the evolution of legal policy surrounding

---

[1] John F. Gantz, *The Expanding Digital Universe*, IDC: ANALYZE THE FUTURE (Mar. 2007), *available at* http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf.

[2] There are several sources about big data volume, management issues, integration issues, etc. The following represent a good survey of extant conversations about Big Data issues: Howard Baldwin, *Big Data's Big Impact Across Industries*, FORBES (Mar. 28, 2014), *available at* http://www.forbes.com/sites/howardbaldwin/2014/03/28/big-datas-big-impact-across-industries/; *See also* Andrew McAfee and Erik Brynjolfsson, *Big Data: The Management Revolution,* HARV. BUS. REV. (Oct. 2012), *available at* http://hbr.org/2012/10/big-data-the-management-revolution/ar; *See also* Thomas H. Davenport, Paul Barth and Randy Bean: *How 'Big Data' Is Different*, MIT BUS. REV. (Jul. 30, 2012), *available at* http://sloanreview.mit.edu/article/how-big-data-is-different/.

[3] Rodney A. Satterwhite & Matthew J. Quatrara, *Asymmetrical Warfare: The Cost of Electronic Discovery in Employment Litigation*, 14 RICH. J.L. & TECH. 9, *available at* http://law.richmond.edu/jolt/v14i3/article9.pdf.

[4] Brian Ingram, *Controlling E-Discovery Costs in a Big Data World*, E-DISCOVERY BRIEF (May 2013), *available at* http://www.lexisnexis.com/legalnewsroom/litigation/b/e-brief/archive/2013/05/02/controlling-e-discovery-costs-in-a-dig-data-world.aspx. Author lays out issues associated with costs, information governance, technology, scope, and efficiency.

scope and size of e-discovery in the age of big data, while arguing that data quality must take precedence in all e-discovery-related endeavors. Finally, it offers productive approaches for securing the most qualitatively valuable, robust dataset in an e-discovery framework, which in turn can be a broader reference for other areas of industry intent on shifting the spotlight from big data to better data.

II. BIG DATA, BIG ESI AND E-DISCOVERY

The digital age has ushered in monumental changes in the discovery process, as the computer and device-driven world in which we live has changed how we generate information from what was once paper-based into a varied electronic landscape.[5] Simply put, the information artifacts of our personal and professional lives are now mostly digital[6] and in the past dozen years or so, ESI has transformed discovery in the litigation process.[7] In 1996, it was estimated that only 5% of discoverable information existed in electronic format.[8] Today, this estimate has increased to over 90%.[9] As ESI grows in capacity with each passing day, as personal and professional communications generate massive amounts of information at accelerated rates, the legal profession has had to cope with the complex issues associated with e-discovery and this information expansion.[10]

To be clear, when we refer to ESI in an e-discovery capacity, we are referring to big data. ESI is a general name that is a catch-all for any and all information produced and housed on a computer or

---

[5] Burke T. Ward, Janice C. Sipior, Jamie P. Hopkins, Carolyn Purwin, Linda Volonino, *Electronic Discovery: Rules for a Digital Age*, 18 B.U. J. SCI. & TECH. L. 150 (2012).

[6] Gantz, *supra* note 1.

[7] *See* Richard L. Marcus, *The Impact of Computers on the Legal Profession: Evolution or Revolution*?, 102 NW. U.L. REV. 4 (2008) (providing a general discussion of issues on how expansion of electronic information has transformed discovery).

[8] Vlad J. Kroll, *Default Production of Electronically Stored Information Under the Federal Rules of Civil Procedure: The Requirements of Rule 34(b)*, 59 HASTINGS L.J. 221, 221 (2007) (stating that in 1996 only 5% of discoverable information came from an electronic format).

[9] David K. Isom, *Electronic Discovery Primer for Judges*, FED. CTS. L. REV. 1, 2 (Feb. 2005), quoting Peter Lyman & Hal R. Varian, *How Much Information*, UNIV. CAL. BERKELEY (2003), *available a*t http://www.sims.berkeley.edu/research/projects/how-much-info-2003/ (92% of all new data is stored and created electronically and 60% of all critical business information is stored within the corporate email system).

[10] Ingram, *supra* note 4.

computing device.[11] Big ESI is big data in the legal profession. Not all ESI is created equal[12] however, and different types of ESI can impact e-discovery processes in different ways.[13] A useful distinction between different types of ESI is structured versus unstructured data. Both data types can be the product of e-discovery, but there are fundamental differences between them.[14] Structured data is that which exists in fixed fields in a file and is easily classified, entered, stored, and queried.[15] It can be text, as entered into fields for example, but it mostly indicates numerical data, or data that has been reduced to a numerical value.[16] Examples include financial information entered and stored in a spreadsheet, click-stream data, input-data such as demographic information stored in a database, and the like. In contrast, unstructured data is that which cannot be readily or automatically classified into neat categories, with tidy, linear correspondences between form and function, or form and meaning.[17] Text-based natural language is unstructured.[18] Examples include social media data like Twitter, blog entries, blog commentary, email, pdf files, PowerPoint presentations, MS Word documents or any file containing natural language.[19]

---

[11] FED. R. CIV. P. 34

[12] Ward et al., *supra* note 5.

[13] Joseph Sremack, *The Collection of Large-Scale Structured Data Systems*, EVIDENCE TECH. MAG., *available at* http://www.evidencemagazine.com/ index.php?option=com_ content&task=view&id=799 (presenting a good overview of the technical differences between identifying and collection structured data collection versus unstructured data).

[14] *Id.*

[15] Vangie Beal, *structured data*, WEBOPEDIA, *available at* http://www.webopedia.com/ TERM/S/structured_data.html (providing a good overview of structured data).

[16] *Id.*

[17] Vangie Beal, *Unstructured Data*, WEBOPEDIA, *available at* http://www.webopedia.com/ TERM/U/unstructured_data.html (providing a good overview of unstructured data); *See also* Cory Jannsen, *Unstructured Data*, TECHOPEDIA, *available at* http:// .techopedia.com / definition /13865/unstructured-data.

[18] Bill Inmon, *Is Text Really Unstructured Data?*, BEYENETWORK, (Mar. 6, 2014), *available at* http://www.b-eye-network.com/view/17247 (providing a good summary of text as unstructured data).

[19] Natural language refers to language that originates from people, as opposed to machine-generated language. *See* Wikipedia, http://en.wikipedia.org/ wiki/ Natural_language.

In today's computer-mediated environment, discoverable ESI exists in the form of vast amounts of unstructured, text-based natural language data.[20] This is especially true in large scale, or complex civil litigation as business communications, as well as a substantial amount of business-related work product, exists as files of unstructured text, in various formats and mediums, created and often stored in unsystematic and uncontrolled manners.[21] Unstructured text-based natural language data qua ESI is, by default, linguistic in nature, although it may contain extra-linguistic information in the form of meta-data that contextualizes it or describes it in a meaningful way.[22] Meta-data is typically defined as "data about data."[23] It can entail information that is attached to a file identifying the source of the file, the date it was created and/or modified, the file type, and so forth.[24] While meta-data can be useful in identifying and describing ESI file characteristics, it does not say anything about the actual contents of the file itself.[25]

Again, the content of unstructured text-based ESI is linguistic in nature, by virtue of the fact that any processes applied to it requires knowledge of language.[26] This bears repeating for a couple of reasons: First, it speaks to the complexity of the data. Natural language and linguistic data are highly dynamic, infinitely variable, innovative, and changes over time.[27] Text-based natural language is no exception.[28]

---

[20] Ward, *supra* note 5, at 9-10.

[21] BARBARA J. ROTHSTEIN, RONALD J. HEDGES, AND ELIZABETH C. WIGGINS, MANAGING DISCOVERY OF ELECTRONIC INFORMATION: A POCKET GUIDE FOR JUDGES 2-3 (2007) (Extrapolated from first and second paragraphs under heading "What Is Electronically Stored Information and How Does It Differ from Conventional Information?").

[22] *See* Wikipedia, http://en.wikipedia.org/wiki/Metadata (describing metadata).

[23] *Id.*

[24] *Id.*

[25] *Id.*

[26] *See generally* DANIEL JURAFSKY AND JAMES H. MARTIN, SPEECH AND LANGUAGE PROCESSING: AN INTRODUCTION TO NATURAL LANGUAGE PROCESSING, COMPUTATIONAL LINGUISTICS, AND SPEECH RECOGNITION 2 (1st ed. 2000) (providing an introduction regarding studying language and language processing as linguistic endeavor).

[27] *See generally* J.K. CHAMBERS, SOCIOLINGUISTIC THEORY: LANGUAGE VARIATION AND ITS SOCIAL SIGNIFICANCE 25-32 (1st ed. 1995) (providing an in-depth discussion on language variation and linguistic systematicity as agents of linguistic processes).

[28] MICHAEL STUBBS, TEXT AND CORPUS ANALYSIS: COMPUTER-ASSISTED STUDY OF LANGUAGE AND CULTURE 33-34 (1st ed. 1996).

Language variation and other linguistic principles characterizing text-based natural language add another layer of complexity to ESI, an already complex unstructured data-type.[29] This is important because it has implications for any empirical research that centers on ideas and content expressed through natural language, but in e-discovery especially, as the basic fact of linguistic variability affects e-discovery processes such as identification, collection and investigation.[30]

Second, the legal profession has historically relied on IT expertise and technology to address matters of handling unstructured text-based ESI in e-discovery processes.[31] Technical expertise is certainly a necessity in e-discovery, but as the nature of ESI is linguistic, it is arguable that language expertise is also very important, especially when approaching qualitative assessments of data.[32] As we will see in the following section, the technical and linguistic complexities of discoverable ESI have another significant component with respect to both data quality and quantity: The legal rules governing it.

III. BIG ESI AND LEGAL POLICY

A. *Background and History*

The advent of Big Data in combination with a legal policy favoring inclusive discovery created a perfect storm for the legal community. This section will discuss how with the guidance of the Federal Rules of Civil Procedure ("FRCP"), the legal profession manages their unique Big Data problem. In many regards, the legal profession serves as a model for other areas of industry, as issues of quality as well as quantity must be addressed, not merely as an academic exercise, but

---

[29] *Id.*

[30] *See generally* GRAEME KENNEDY, AN INTRODUCTION TO CORPUS LINGUISTICS 62 (1st ed. 1998) (discussing representativeness and balance in corpus design in section 2.5.5 of Chapter 2).

[31] RELIANCE ON IT: THE STATE BAR OF CALIFORNIA STANDING COMM. ON PROF'L RESPONSIBILITY AND CONDUCT FORMAL OP. INTERIM NO. 11-0004 ("An attorney lacking the required competence for the e-discovery issues in the case at issue has three options: (1) acquire sufficient learning and skill before performance is required; (2) associate with or consult technical consultants or competent counsel; or (3) decline the client representation"); *See also* "Gartner Predicts Growth, Consolidation in E-Discovery Market," ARMA INT'L (Jul. 24, 2013), *available at* http://www.arma.org/r1/news/newswire/2013/07/24/gartner-predicts-growth-consolidation-in-e-discovery-market (discussing the growth of IT based legal discovery).

[32] Stubbs, *supra* note 28, at 51-52.

to sever the very serious interests of justice. Before beginning a substantive discussion of Big Data and legal policy, however, here is an overview of the FRCP, including what is meant by discovery, relevance, and the "liberal policy of inclusion."

The FRCP is the set of regulations that specify procedures for civil legal suits within the United States District Courts.[33] Federal district courts in all fifty states are required to follow these rules, and many state courts' civil procedural rules closely follow or adopt similarly worded rules.[34] "These...should be construed and administered to secure the just, speedy and inexpensive determination of every action and proceeding."[35] Title V of the FRCP contains the rules that specifically address disclosure and discovery.[36]

Legal discovery is the process of uncovering *relevant* facts through identifying witnesses, documents and other items that can lead to establishing those facts as admissible evidence. It is the formal process of exchanging information between the parties about the witnesses and evidence they will present at trial.[37] Discovery enables the parties to know before the trial begins what evidence may be presented. It is designed to prevent "trial by ambush," where one side does not learn of the other side's evidence or witnesses until the trial, when there is not time to obtain answering evidence.[38] While establishing facts in the discovery process can take on many forms, such as depositions, both oral and written[39], interrogatories[40] or even the physical or mental examination of a party[41], the discussion of this

---

33 Eric Hibbard, *How the Latest FRCP Changes Should Put Experts on Notice*, HITACHI DATA SYS. (Jan. 19, 2011), *available at* http://blogs.hds.com/hdsblog/2011/01/how-the-latest-frcp-changes-should-put-experts-on-notice.html.

34 *Id.*

35 FED. R. CIV. P. 1 (addressing scope and purpose).

36 FED. R. CIV. P. Title V (addressing disclosure and discovery).

37 American Bar Association, Division for Public Education, *How Courts Work*, *available at* http://www.americanbar.org/ groups/public_education/resources/ law_related_education_ network/ how_courts_work/discovery.html.

38 *Id.*

39 FED. R. CIV. P. 30. (Depositions by Oral Examination); FED. R. CIV. P. 31. (Depositions by Written Questions).

40 FED. R. CIV. P. 33 (Interrogatories to Parties)

41 FED. R. CIV. P. 35 (Physical and Mental Examinations).

paper focuses strictly on the discovery of documents, particularly those known as ESI.[42]

Currently, relevancy is broadly defined under the FRCP. Rule 26(b)(1) specifies that, "unless otherwise limited by court order . . .":

> Parties may obtain discovery regarding **any** non-privileged matter that is relevant to any party's claim or defense, including the existence, description, nature, custody, condition, and location of any documents or other tangible things and the identity and location of persons who know of any discoverable matter. For good cause, the court may order discovery of any matter relevant to the subject matter involved in the action. Relevant information need not be admissible at the trial if the discovery appears reasonably calculated to lead to the discovery of admissible evidence. (emphasis added)

Discovery is not limitless, however. All discovery is subject to the limitations imposed by Rule 26(b)(2)(C).[43] Indeed, as early as 1947, the U.S. Supreme Court cautioned that discovery has "ultimate and necessary boundaries" that include inquiries into irrelevant or privileged matters or those conducted in bad faith.[44] Additional specific limits apply to ESI[45] and will be discussed in detail below.

---

[42] FED. R. CIV. P. 34 (Producing Documents, Electronically Stored Information, and Tangible Things, or Entering onto Land, for Inspection and Other Purposes).

[43] FED. R. CIV. P 26(b)(2)(C) ("When Required. On motion or on its own, the court must limit the frequency or extent of discovery otherwise allowed by these rules or by local rule if it determines that: (i) the discovery sought is unreasonably cumulative or duplicative, or can be obtained from some other source that is more convenient, less burdensome, or less expensive; (ii) the party seeking discovery has had ample opportunity to obtain the information by discovery in the action; or (iii) the burden or expense of the proposed discovery outweighs its likely benefit, considering the needs of the case, the amount in controversy, the parties' resources, the importance of the issues at stake in the action, and the importance of the discovery in resolving the issues.").

[44] Hickman v. Taylor, 329 U.S. 495, 507-08 (1947).

[45] FED. R. CIV. P. 26(b)(2)(B) ("Specific Limitations on Electronically Stored Information. A party need not provide discovery of electronically stored information from sources that the party identifies as not reasonably accessible because of undue burden or cost. On motion to compel discovery or for a protective order, the party from whom discovery is sought must show that the information is not reasonably accessible because of undue burden or cost. If that showing is made, the court may nonetheless order discovery from such sources if the requesting party shows good cause, considering the limitations of Rule 26(b)(2)(C). The court may specify conditions for the discovery").

B. *A Big Data Explosion: Law Acts…and Then Reacts*

Since their inception in 1938, although the FRCP have had to evolve to accommodate a changing discovery landscape,[46] the foundational aspects of discovery remain constant[47] and liberal discovery remains the norm in civil litigation.[48] The most recent change in the discovery landscape, that which is most pertinent to a discussion of Big Data, is the 2006 Amendments to the FRCP. The 2006 Amendments both officially confirmed that ESI stands on equal footing with discovery of paper documents and is thus subject to the same discovery laws as traditional paper or tangible documents[49] and recognized that ESI presents a different set of discovery challenges,[50] one of which is, unsurprisingly, volume, or quantity.

---

[46] The original Federal Rules of Civil Procedure for the District Courts were adopted by order of the Supreme Court on Dec. 20, 1937, transmitted to Congress by the Attorney General on Jan. 3, 1938, and became effective on Sept. 16, 1938. Significant revisions have been made to the Rules in 1948, 1963, 1966, 1970, 1980, 1983, 1987, 1993, 2000, and 2006. *See* FED. R. CIV. P Historical Note for a full list of amendments and dates the amendments were effected, available at http://www.law.cornell.edu/rules/frcp.

[47] The Court has reiterated this broad standard, stating that it "has more than once declared that the deposition-discovery rules are to be accorded a broad and liberal treatment to effect their purpose of adequately informing the litigants in civil trials. Herbert v. Lando, 441 U.S. 153, 177 (1979) (citing Schlagenhauf v. Holder, 379 U.S. 104, 114-115 (1964); Hickman, 329 U.S. at 501); See also Sanyo Laser Products, Incorporated v. Arista Records, Incorporated, 214 F.R.D. 496, 500 (S.D. Ind. 2003); Accord Fountain v. City of New York, No. 03 Civ. 4526 (S.D.N.Y. May 3, 2004); See also Henderson v. Property and Casualty Insurance Company of Hartford, No. 2:12-cv-00149 (D. Nev. Aug. 28, 2012) ("Most courts which have addressed the issue find that . . . Rule 26 still contemplate[s] liberal discovery, and that relevancy under Rule 26 is extremely broad."); Wrangen v. Pennsylvania Lumbermans Mutual Insurance Company, 593 F. Supp.2d 1273, 1278 (S.D. Fla. 2008) ("[D]iscovery should ordinarily be allowed under the concept of relevancy unless it is clear that the information sought has no possible bearing on the claims and defenses of the parties or otherwise on the subject matter of the action." (quoting Dunkin' Donuts, Inc. v. Mary's Donuts, Inc., No. 01-0392-Civ-Gold, 2001 WL 34079319 *2 (S.D.Fla. Nov. 1, 2001)).

[48] *Id.*

[49] 2006 Amendments to FED. R. CIV. P. 34, advisory committee's note ("Rule 34(a) is amended to confirm that electronically stored information stands on equal footing with discovery of paper documents.").

[50] "In addition to its sheer volume, electronically stored information (ESI) is also distinguished from tangible documents by its complexity and availability in increasingly diverse formats." *Federal Practice Manual for Legal Aid Attorneys*, SARGENT SHRIVER NAT'L CENTER ON POVERTY LAW, 6.2.E. Electronic Discovery, *available at* http://federalpracticemanual.org/node/34.

As the universe of digital data was expanding,[51] so was the volume of discoverable ESI. To put this in perspective, researchers at IDC, a global marketing intelligence firm, released a white paper in 2007 that determined the amount of digital information created and replicated the previous year was equivalent to about 3 million times the volume of books ever written up to that point in history.[52] Furthermore, the research suggested that 70% of all digital information was created by individuals and organizations of all sizes and caliber, from small companies to government agencies.[53] Liberal discovery policy and the rapid growth of ESI meant that legal teams routinely found themselves in an e-discovery torrent with potentially enormous amounts of relevant ESI that was to become the centerpiece for fact-finding and intelligence gathering to support, or refute, legal narratives associated with a case. What's more, legal teams had to devise and implement discovery processes for identifying, preserving, collecting, reviewing, producing and investigating potentially enormous amounts of relevant ESI in a universe of even greater amounts of disparate, irrelevant data.

Under these circumstances, the fixation on quantity is understandable. It is hard to talk about e-discovery without talking mostly about volume and the cost that accompanies it.[54] It is also

---

[51] In 2000, and again in 2003, before the phrase Big Data was coined, researchers at University of California at Berkeley conducted a study to determine "How much information is produced each year?" Even though this study is more than a decade old, highlights from the study's executive summary report are still impressive: "Print, film, magnetic, and optical storage media produced about 5 exabytes of new information in 2002. Ninety-two percent of the new information was stored on magnetic media, mostly in hard disks. Hard disks store most new information. Ninety-two percent of new information is stored on magnetic media, primarily hard disks. Film represents 7% of the total, paper 0.01%, and optical media 0.002%. Instant messaging generates five billion messages a day (750GB), or 274 Terabytes a year. Email generates about 400,000 terabytes of new information each year worldwide. Peter Lyman, Hal R. Varian, *How Much Information* (2000), *available at* http://www.sims.berkeley.edu/how-much-info.

[52] Gantz, *supra* note 1 (5th bullet point under key findings).

[53] *Id.*

[54] In the period since 2006, e-discovery processes have been the fastest growing areas of cost and expense associated with civil litigation, out-pacing other legal fees. *See* George Socha, Tom Gelbmann, *Mining for Gold*, LAW TECH. NEWS, Aug. 5, 2008. In 2007, for example, litigants spent nearly $2.79 billion dollars on e-discovery, a 43% increase from the amount spent just a year earlier. *See also* Nicholas Pace, Laura Zakaras, *Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery*, RAND INST. FOR CIVIL JUSTICE, (2012), a case study of Fortune 500 companies, finds that the median total cost for ESI production among participants reached the sum of $1.8 million dollars per case.

difficult to discuss identifying, preserving and producing relevant ESI without considering the vast universe and vast quantities in which that potentially relevant data reside. In addition, with affirmative obligations now squarely resting upon counsel to understand and manage this brave new world of ESI[55] as well as not unfounded fears that even in the absence of willfulness or bad faith, a court may still impose sanctions[56], the situation was ripe for discovery to become a protracted, expensive, ugly battle of attrition. While accusations of "abuse of discovery" are nothing new in our adversarial system[57], the post-2006 discovery environment is rife with accusations of "hide the ball tactics"[58] fishing expeditions[59], and demands for court imposed sanctions.[60] Even worse, the threat of a protracted and/or expensive discovery process lead to "forced settlements"[61] simply because the parties could not afford (in terms of sheer cost, or in terms of a production timeframe that could take years to complete) to see the matter tried on its merits. Whether you interpret "discovery abuse" as

---

[55] *See generally* Qualcomm, Inc. v. Broadcom Corp., 539 F. Supp. 2d 1214 (S.D. Cal. 2007).

[56] *See* Pension Comm. of the Univ. of Montreal Pension Plan v. Banc of America Securities, LLC, 685 F. Supp. 2d 456 (S.D.N.Y. 2010) (plaintiffs' conduct described as "either grossly negligent or negligent because they failed to execute a comprehensive search for documents and/or failed to sufficiently supervise or monitor their employees' document collection); *See also* Tony Schoenberg, "Steps for Avoiding Sanctions," *California Lawyer A Daily Journal Publication* October 2010, *available at* http:/ /www.callawyer.com /Clstory.cfm?eid=911880 (discussing how level of culpability required for imposing sanctions can vary from court to court).

[57] *See* Ralph Losey, *Fears and Loathing (and Pain) in Seattle: a Case Lesson in How NOT to Implement a Litigation Hold and Search for Email – Part Two, e-discovery team*, April 20, 2014, available at http://e-discoveryteam.com/category/review/ (discussing and listing sites related to discovery abuse).

[58] *See* Ralph Losey, "*Fears and Loathing (and Pain) in Seattle: a Case Lesson in How NOT to Implement a Litigation Hold and Search for Email – Part One, e-discovery team*," April 13, 2014, available at http://e-discoveryteam.com/page/7/.

[59] Losey, supra note 57.

[60] *Id.*

[61] Fed. R. Civ. P. Notes of Advisory Committee on Rules—1983 Amendment: Excessive discovery and evasion or resistance to reasonable discovery requests pose significant problems. Recent studies have made some attempt to determine the sources and extent of the difficulties. *See* BRAZIL, CIVIL DISCOVERY: LAWYERS' VIEWS OF ITS EFFECTIVENESS, PRINCIPAL PROBLEMS AND ABUSES (1980); CONNOLLY, HOLLEMAN & KUHLMAN, JUDICIAL CONTROLS AND THE CIVIL LITIGATIVE PROCESS: DISCOVERY (1978); Schroeder & Frank, The Proposed Changes in the Discovery Rules, 1978 ARIZ. ST .L.J. 475.

the defendants in your cases "refusing to supply information"[62] or "as a lever to force settlement in cases that have little merit,"[63] it is clear that there was growing dissatisfaction with the FRCP, particularly with discovery and litigation costs.[64] But with the "new" Federal Rules not even a decade old, would there be any action?

C. *The Proposed "New" Rules*

In August 2013, after much study, the Judiciary Conference's Advisory Committee on Civil Rules ("Advisory Committee") proposed further amendments to the FRCP that are likely to go into effect in December of 2015. In June 2014, the Advisory Committee reviewed the public comments and hearing testimony and has recommended the rule changes to the Judicial Conference. The Judicial Conference met in September 2014, where it recommended the revisions to the U.S. Supreme Court, which may then promulgate any revisions on or before May 15, 2015.[65] Any such revision that is implemented will take effect on December 1, 2015, unless Congress rejects, modifies or defers

---

[62] EMERY G LEE III, THOMAS E WILLGING, ATTORNEY SATISFACTION WITH THE FEDERAL RULES OF CIVIL PROCEDURE, REPORT TO THE JUDICIAL CONFERENCE ADVISORY COMMITTEE ON CIVIL RULES, n. 13 (Mar. 2010).One [survey] respondent commented, "Discovery abuse is rampant – parties (usually defendants) stonewall routinely and then negotiate over how many of their legal obligations they can avoid." Another commented that costs would be reduced if judges would "[e]nforce sanctions for discovery abuses. Much of the costs we deal with relate to trying to get sufficient discovery – the delay and the cost of filing motions to compel, etc., increase costs significantly."

[63] *Id.* at n.14. One ABA Section defendant attorney commented, "Demands for e-discovery are being used as a lever to force settlement in cases that have little merit. Most e-discovery is useless and should not be requested in the first instance."

[64] *Id.* at 3. "The Advisory Committee on Civil Rules ("Committee") requested that the Federal Judicial Center study, among other things, whether attorneys are generally satisfied with the present operation of the Federal Rules of Civil Procedure. This request followed a joint report issued by the American College of Trial Lawyers and the Institute for the Advancement of the American Legal System ("'IAALS'"), based on a survey of ACTL fellows. In summarizing the survey results, the ACTL/AALS joint report stated: "'In short, the survey revealed widely-held opinions that there are serious problems in the civil justice system generally.'""

[65] *Summary Of The Report Of The Judicial Conference Committee On Rules Of Practice And Procedure*, Agenda E-19 (Summary) Rules September 2014, at 13, *available at* http://www.uscourts.gov/uscourts/RulesAndPolicies/rules/Reports/ST09-2014.pdf; Nash E. Long, *The Proposed Amendment to the Federal Rules of Civil Procedure on Discovery*, ABA, April 10, 2014, *avaiable at* http://www.americanbar.org/c ontent/dam/aba/ administrative/ litigation/materials/ 2014_sac/ 2014_sac/ the_proposed _amendments. authcheckdam.pdf.

the same.[66] These amendments are thought to be most significant changes to discovery in the past 25 years.[67]

The proposed rules, "among other things, would narrow the scope and require proportionality of discovery . . ."[68] The proposed changes to Rule 26(b)(1) significantly alters current practice. First, the revised language incorporates the concept of proportionality as a condition of entitlement to discovery. The proposed rule provides certain factors a court should consider in determining proportionality. These include the amount in controversy, importance of issue at stake, parties' resources, importance of discovery to resolve an issue, and expense versus the benefit.[69] Second, the proposed amendment makes a very significant change in terms of the permissible scope of discovery by removing the language from the current rule which states: "Relevant information need not be admissible at the trial if the discovery appears reasonably calculated to lead to the discovery of admissible evidence." This means that under the proposed rule, the standard is a pure relevancy standard, and nothing more.[70]

The proposed amendments represent a palpable shift in the courts approach to managing e-discovery, and will create an opportunity for the legal profession to refine its approach to e-discovery as well. There is an opportunity for qualitatively robust discovery to become the rule, rather than the exception in civil litigation. It is understandable that issues of quantity and scope have dominated e-discovery for nearly a decade: All industries, from health care to horse racing, have been trying to come to terms with big data.[71] However, the persistent focus on quantity and scope of ESI in e-discovery in particular, and the legal profession in general, has meant that conversations regarding quality

---

[66] *Id.*

[67] Marc A. Goldich, David R. Cohen, Emily J. Dimond*, FRCP Amendments Could Change Discovery As We Know* It, LAW 360 (June 2013), *available at* http://www.law360.com/articles/447209/frcp-amendments-could-change-discovery-as-we-know-it.

[68] *Id.*

[69] Long, *supra* note 65.

[70] *Id.*

[71] Alex Frommeyer, *Can Big Data Save Horse-Racing Industry?*, LOUISVILLE BIZBLOG (May 5, 2014), *available at* http://www.bizjournals.com/louisville/blog/2014/05/can-big-data-save-horse-racing-industry.html?page=all; See also Jordan Robertson, *The Health-Care Industry Turns to Big Data*, BLOOMBERG BUSINESSWEEK (May 17, 2013), available at http://www.businessweek.com/articles/2012-05-17/the-health-care-industry-turns-to-big-data. These are just two examples of many that turn up on a search for a big data and industry/business.

have been all but absent.[72] ESI quality will matter more than ever in e-discovery, as protocols and processes will have to focus on identifying and producing *highly relevant* documents, the most qualitatively valuable datasets in order to comply with proportionality tests while continuing to be administered to secure the just, speedy, and inexpensive determination of every action and proceeding. In terms of big civil litigation, it is still a big ESI universe, but it is now going to be necessary to explore, define and implement qualitative parameters to that universe. These exercises in quality are overdue, but also timely in the face of evolving legal policy regarding e-discovery.

IV. THE BEST ESI

A. Relevancy and Data Quality

Now it is time to explore what quality means in the context of ESI and e-discovery processes *sans* references to quantity or scope; quality for quality's sake, as it were. First and foremost, talking about qualitatively valuable ESI in the context of e-discovery requires an expanded discussion about relevancy. We have already examined the legal concept of relevancy as a foundation for discovery processes. Identifying a document or an email, or a phrase in an email or some meta-data describing a document, as meeting some established threshold for relevancy is a legal exercise to be sure, but it also speaks to the process of examining the language of ESI itself, and therefore it is an empirical exercise in data quality as well.[73] A general way to look at it is that a qualitative concept of relevancy must have purchase in real world phenomena.[74] Specifically, a qualitative discussion about

---

[72] *The Sedona Conference, The Sedona Conference Commentary on Achieving Quality in the E-Discovery Process: A Project of The Sedona Conference Working Group on Electronic Document Retention & Production*, WG1 (Dec. 2013), *available at* https://thesedonaconference.org/publication/The%20Sedona%20Conference%C2%AE%20Commentary%20on%20Achieving%20Quality%20in%20the%20E-Discovery%20Process. Most of what is written on "quality" and "quality control" and "quality assurance" exists as company marketing pieces, company white papers, e-discovery and legal blog entries, and the like. This is an opinion piece put out by the Sedona Conference that directly addresses the need for quality in e-discovery.

[73] *Id.* at 15 (discussing relevancy identification, especially in "3 *When are Quality Measures and Metrics Appropriate?"); See also* footnote on page 15 (identifying relevant material in terms of quality).

[74] *See generally* MICHAEL STUBBS, WORDS AND PHRASES: CORPUS STUDIES OF LEXICAL SEMANTICS 1-9 (2001) (discussing, in chapter 1, about how we map semantic meaning and /cultural meaning onto words and phrases, drawing from our ability to imbue a form with an entire knowledge complex, in turn giving us the ability to assign order and value to the world around us).

relevancy needs to include how relevant information is concretely manifested in ESI as unstructured, text-based natural language data, as this is the common data type in e-discovery.[75] It is therefore reasonable to presuppose that when it comes to discoverable ESI, data quality entails linguistic quality.[76]

At its most basic level, relevant ideas, concepts and facts are expressed through language usage: Indeed, it is generally accepted that language is a medium for the communication of any idea or concept.[77] As such, ideas or concepts of what is qualitatively relevant information cannot be divorced from the text-based medium or form used to express it.[78] Thus, the language itself, the text, *is* the potentially relevant evidence.[79] It is reasonable to assume then that language as potential evidence requires acknowledgment and accommodation of basic linguistic principles that are in play when identifying and attaching relevant meaning to a word, a phrase, or a chunk of text.[80]

The first linguistic principle that affects the identification and description of qualitatively relevant information in text-based ESI is language variation. Language is infinitely variable.[81] When it comes to language, even text-based language, there are a multitude of ways, or a variety of linguistic forms, in which to express an idea or a fact.[82] Thus, if you are operating at a conceptual level in assigning relevant narratives or themes in a case during the e-discovery planning stage, know that there are variable ways to express the same concepts in any given language. Variation is simply an empirical reality of natural

---

[75] *See* Ward et al., *supra* note 5, at 9.

[76] *See generally* MICHAEL STUBBS, TEXT AND CORPUS ANALYSIS: COMPUTER-ASSISTED STUDY OF LANGUAGE AND CULTURE (1st ed. 1996) (providing a comprehensive examination and history of linguistics and text analysis). The idea of text analysis as a linguistic endeavor is based on the idea that research, involving extracting meaning or knowledge from text, has long-standing empirical traditions in linguistic theory and practice.

[77] *See, e.g.*, any definition of "language", *available at* http://www.oxforddictionaries .com/us/ definition/american_english/language.

[78] *See generally* Stubbs, *supra*, at 28.

[79] *Id.*

[80] *Id.* at 35.

[81] *See generally* PETER SIEMUND, LINGUISTIC UNIVERSALS AND LANGUAGE VARIATION 1-22 (2011).

[82] JOHN SINCLAIR, CORPUS, CONCORDANCE, COLLOCATION 4 (1991).

language use.[83] A quick illustration of this involves a small corpus linguistic experiment[84] conducted to look at lexio-semantic units[85] as associated with the meaning of *hit*.[86] The researcher found that the term had a wide range of uses, one of which was to talk about traffic accidents. However, there were a variety of other lexio-semantic units associated with the traffic accident concept within the same semantic field as *hit*,[87] terms such as bumped, smashed, collided, and struck.[88] As such, you could extrapolate a variety of sentences that all expressed roughly the same fact: His car hit mine. His car bumped into mine. His car smashed into mine. His car collided with mine. His car struck mine.

While language is infinitely variable, language is also a habit and luckily, another empirical reality of language is that in terms of linguistic features like lexio-semantic units, or linguistic types, there are linguistic norms that describe frequency of usage.[89] You can quantitatively derive a linguistic norm that offers visibility into what particular linguistic variety is the most common versus what variety is the most rare. You can use linguistic norms generally to understand what language is disproportionately frequent within a corpus of text-based natural language or what is disproportionately infrequent.[90]

Another important principle that will impact relevancy as a qualitative exercise is that meaning is derived from context, and context can be both linguistic and extra-linguistic.[91] There is a famous

---

[83] Siemund, *supra* note 81.

[84] Corpus linguistics focuses written or transcribed text as the foundation of linguistic analysis and description. *See generally* GRAEME KENNEDY, AN INTRODUCTION TO CORPUS LINGUISTICS (1st ed. 1998).

[85] A lexio-semantic unit is roughly equivalent to an open-class word that has meaning attached to it, as opposed to say a closed-class word that is used for grammatical purposes, such as "would" or "have."

[86] MICHAEL STUBBS, WORDS AND PHRASES: CORPUS STUDIES OF LEXICAL SEMANTICS 118-119 (2001).

[87] Semantic field is roughly the same thing as a general concept qua word that has meaning attached to it, like “ball” or “run,” for example.

[88] Stubbs, *supra* note 76.

[89] *See generally* MICHAEL P. OAKES, STATISTICS FOR CORPUS LINGUISTICS 3-4 (1998).

[90] Oakes, *supra* note 89.

[91] MICHAEL STUBBS, TEXT AND CORPUS ANALYSIS: COMPUTER-ASSISTED STUDY OF LANGUAGE AND CULTURE 53 (1st ed. 1996). Any reference to social influences on context, or any influence that is not linguistic in nature, is referred to in the field as extra-linguistic.

and oft-cited saying in empirical linguistics that perfectly describes linguistic context: Words shall be known by the company they keep.[92] Referring back to the previous example used to illustrate language variation, let's again consider the semantic field *hit* in the following sentence*:* His car was hit by lightning. Other semantic elements of "traffic accident" are there, but this sentence is clearly not referring to a traffic accident. It is necessary to look at the linguistic context *in toto* in order derive meaning from it. Extra-linguistic factors are also at play in deriving meaning: As far as relevancy and meaning are concerned, who is saying something, why it is being said, as well as when it is being said, is just as important as what the linguistic context is conveying.[93] Again, going with our example, consider a police report describing an officer pulling over a driver seen fleeing the scene of an accident in a badly damaged car. The driver fails a sobriety test. The officer asks the driver what happened to his vehicle and the driver replies that it was struck by lightning, prompting the officer to dictate in the report: The car was struck by lightning. In this scenario, all of the linguistic and extra-linguistic factors are necessary to properly contextualize and interpret meaning from the text.

These very elementary examples illustrate very complex but universal linguistic phenomena that impact language as potential evidence with respect to identifying and assigning relevancy. Let's put these linguistic principles in the context of an e-discovery and civil litigation scenario. Consider a typical products liability case involving a pharmaceutical company's DRUG X. Safety has been identified as a relevant concept. First, how is safety expressed in the universe of ESI in play? What is the language variation associated with the concept? Are all instances of safety in the corpus relevant? What about the lemma of the term safety: Safe, safer, safest, safeties, safely, unsafe, safeguard? Do all of these impart safety as a concept? What about language that does not include any of these lemma, but rather talks about possible harm (or no possible harm) to those who are administered Drug X? Is an email that states as much considered relevant to the concept of safety? Language variation aside, is an email from a 3rd party adverse event reporting site talking about possible harmful side effects as relevant as a memo from the chief scientist of product development to the company's CEO talking about possible harmful side effects?

---

92 JACQUELINE LÉON, *MEANING BY COLLOCATION. THE FIRTHIAN FILIATION OF CORPUS LINGUISTICS*, PROC. OF ICHOLS X, 10TH INT'L CONF. ON THE HIST. OF LANGUAGE SCI. 404 (2007).

93 NELLEKE OOSTDIJK, CORPUS LINGUISTICS AND THE AUTOMATIC ANALYSIS OF ENGLISH 52 (2011) (discussing the range of extra-linguistic variables).

When language is potentially evidence, the decision-making processes used to identify and determine what has relevant significance or meaning will have to take these empirical facts about language into consideration in order to ensure a qualitatively robust dataset. If language variation is not recognized, investigated, or generally taken into account, as well as contextualizing information that informs meaning, it will be difficult to ensure that relevant ESI is not overlooked and therefore, not disclosed. Even if advanced technologies and methodologies are used for filtering a production for relevance, it is still necessary to understand and account for the facts of language as potential evidence that will influence these processes. Linguistic knowledge is crucial not only in working with text-based natural language data, but in understanding the limits and benefits of *any* methodology used to analyze ESI with respect to relevancy, or any empirical research endeavor.[94]

Moreover, relevancy is not the only qualitative issue in e-discovery that can benefit from linguistic knowledge. Like relevancy, the identification of potentially privileged communications is an issue of quality.[95] There are legal tests for privilege that include, but are not limited to, whether the communication is between an attorney and a client, and whether or not the client is seeking a legal opinion, advice or assistance in the context of the communication.[96] Thus there are both linguistic and extra-linguistic factors at play in identifying and designating a communication as potentially privileged, just as there are in identifying a communication as potentially relevant. Language expertise and linguistic-based processes are poised to assist in all issues of quality in an e-discovery capacity, from pre-production identification and collection, to post-production investigation and analysis.

B. *Attaining Quality in Big ESI*

Best practices with respect to issues of quality in e-discovery will likely change as legal policy evolves and data-driven technology advances. Legal policy and technology are certainly key components of the quality equation in e-discovery. However, they are not the only important aspects of achieving quality in a digital era preoccupied with quantity. Ensuring ESI quality in e-discovery requires a

---

[94] *See generally* The Sedona Conference, *supra* note 72, at 15 (providing a discussion regarding experts and non-experts employing tools for measuring quality).

[95] The Sedona Conference, *supra* note 72, at 30.

[96] FED. R. EVID. 502.

marriage of legal expertise, technical expertise, general subject-matter expertise and linguistic expertise. In the context of a large civil suit, where a substantial discovery effort is anticipated, the ideal team assembled to plan and execute a solid e-discovery strategy would include all of these areas of expertise. The legal experts would have a-priori, case-specific knowledge that would serve as the framework for developing an e-discovery plan that balances quality with quantity. The technical experts, in both digital forensics and information technology, would oversee the data gathering and data transformation processes. The linguistic expert would provide scientific, principled, data-driven methodologies and insight to processes of identifying, managing, and investigating unstructured, text-based natural language. Subject matter experts would provide industry or professional knowledge needed to make industry-specific information accessible to non-experts. Having access to the right combination of expertise is the most important aspect of ensuring a qualitatively robust dataset, not just in e-discovery, but in any context in which unstructured text-based natural language data is the centerpiece for empirical research. Ultimately, it is the combination of these areas of expertise that is the only way that issues of quality versus quantity can be resolved in a reasonable, valid, reliable and defensible manner.

V. LOOKING AHEAD

Those working with ESI in e-discovery in the legal sector face challenging, yet exciting opportunities to address issues of data quality along with issues of data quantity. The legal profession has been decisive in its efforts to deal with the practical matters that come with big data territory. Legal experts have examined and reexamined their policies regarding big ESI and discovery and are positioned to realign industry standards in trying to achieve quality-quantity balance in data-driven, empirical research in the age of big data. If we can learn anything about the big data revolution from the legal profession, it's that bigger doesn't necessarily mean better.